DOCUMENT RESUME

ED 409 357 TM 026 880

AUTHOR Scrams, David J.; Schnipke, Deborah L.
TITLE Making Use of Response Times in Standardized Tests: Are Accuracy and Speed Measuring the Same Thing?
PUB DATE Mar 97
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Assisted Testing; Difficulty Level; Item Response Theory; Psychometrics; *Reaction Time; *Standardized Tests; *Test Items; Thinking Skills; *Timed Tests
IDENTIFIERS *Accuracy; Large Scale Assessment; *Speededness (Tests)

ABSTRACT

Response accuracy and response speed provide separate measures of performance. Psychometricians have tended to focus on accuracy with the goal of characterizing examinees on the basis of their ability to respond correctly to items from a given content domain. With the advent of computerized testing, response times can now be recorded unobtrusively during operational tests, and this new source of data may provide additional information about examinees. D. Thissen (1983) offered an extension of item response theory that accounts for both accuracy and speed within a single model. Thissen's Timed-Testing model is used in this study as a framework for exploring the relationship between accuracy and speed in three large-scale computerized tests. Data are from computer-administered tests of verbal, quantitative, and reasoning skills involving about 7,000 examinees. Overall relative item easiness accounted for only a small proportion of variability in response times, and neither examinee ability nor item difficulty performed much better. Results are discussed in terms of speededness and the possibility of incorporating speed factors into ability estimation. (Contains 5 figures and 11 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 409 357

Making Use of Response Times in Standardized Tests: Are Accuracy and Speed Measuring the Same Thing?

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
David J. Scrams
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

David J. Scrams
Deborah L. Schnipke

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM026880

This paper is prepared for the Annual Meeting of the American Educational Research Association in Chicago, IL

BEST COPY AVAILABLE

Paper presented at the Annual Meeting of the
National Council on Measurement in Education
Chicago, IL March, 1997

Making Use of Response Times in Standardized Tests: Are Accuracy and Speed Measuring the Same Thing?

David J. Scrams

Johns Hopkins University

Deborah L. Schnipke

Law School Admission Council

Response accuracy and response speed provide separate measures of performance. Psychometricians have tended to focus on accuracy with the goal of characterizing examinees on the basis of their ability to respond correctly to items from a given content domain. With the advent of computerized testing, response times can now be recorded unobtrusively during operational tests, and this new source of data may provide additional information about examinees. Thissen (1983) offered an extension of Item Response Theory that accounts for both accuracy and speed within a single model. Thissen's Timed-Testing model is used in this study as a framework for exploring the relationship between accuracy and speed in three large-scale, computerized tests. Results are discussed in terms of speededness and the possibility of incorporating speed factors into ability estimation.

Psychometricians and cognitive psychologists share an interest in human performance, but their goals and approaches are often widely divergent. Psychometricians seek to explain performance in terms of examinee and item characteristics, and cognitive psychologists would prefer to focus on cognitive processes and operations.

These approaches are not incompatible, and some test theorists have strongly advocated increased integration of psychometric and cognitive theory (Embretson, 1983, 1985; Mislevy, 1994). Although there are a number of research areas that could benefit from such integration, psychometricians investigating response times may find the cognitive literature a particularly useful resource.

Response Times in the Testing Context

Classical Test Theory (CTT) and Item Response Theory (IRT) focus exclusively on response accuracy; examinee performance is measured in terms of the number (CTT) or characteristics (IRT) of items answered correctly and incorrectly. The exclusion of response-time information from psychometric models is understandable given that response times cannot be recorded easily during traditional paper-and-pencil test administrations. This limitation has been eliminated by computerized testing, so response times can now be recorded easily and unobtrusively during operational tests.

The sudden availability of item response times suggests a wide range of opportunities for psychometric theorists. Although response times were considered in the testing context at least 60 years ago (Thurstone, 1937), the theoretical issues involved have been largely ignored because response times were simply not available. Psychometricians rushing to fill this gap have found themselves in the potentially exciting but often frustrating position of having little previous work on which to rely. Luckily, although psychometric treatment of response-time data has been relatively minimal, response time has been the predominant dependent variable used by cognitive psychologists (Luce, 1986).

The most rigorous treatment of response time in cognitive psychology has been in the context of mathematical models of information processing, and the most powerful of these models account for both response accuracy and response time simultaneously (Townsend & Ashby, 1983). Thus, much of cognitive psychology has been devoted to the relationship between accuracy and speed in various contexts. This is an important issue in the testing context as well.

The Present Work

The present work draws on cognitive and psychometric theory to explore the relationship between accuracy and speed in the testing context. First, two different types of speed-accuracy relationships are discussed: a within-examinee relationship (the "speed-accuracy tradeoff" of cognitive psychology) and an across-examinee relationship (perhaps of more interest to psychometricians). The remainder of the

We would like to thank Bert Green, Louis Roussos, and three anonymous reviewers for helpful comments during the initial stages of this project.

Correspondence concerning this manuscript should be addressed to David J. Scrams, Psychology Department, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218.

manuscript focuses on the second type of speed-accuracy relationship.

An integrative model proposed by Thissen (1983), the Timed-Testing model, is discussed. This model is an extension of Item Response Theory that accounts for both accuracy and speed. The model is applied to data from three large-scale, computer-administered tests, and results are discussed in terms of test speededness and the possibility of incorporating response-time information into ability estimation.

A Cognitive Psychology Approach to Speed-Accuracy Relationships

When cognitive psychologists investigate speed-accuracy relationships, they generally focus on within-subject effects. They address questions such as "How does task performance relate to processing time?" By forcing subjects to respond at given deadlines and varying these deadlines on a within-subject basis while subjects respond to a large number of interchangeable items, cognitive psychologists can sweep out a function relating response accuracy to response speed for a given individual performing a given task (Dosher, 1981; Reed, 1973). Multiple subjects are used so that common characteristics emerge, and these so-called *speed-accuracy tradeoff functions* are compared across different tasks with the goal of uncovering the cognitive processes or operations that underlie task performance. Individual differences in these functions are interpreted as noise.

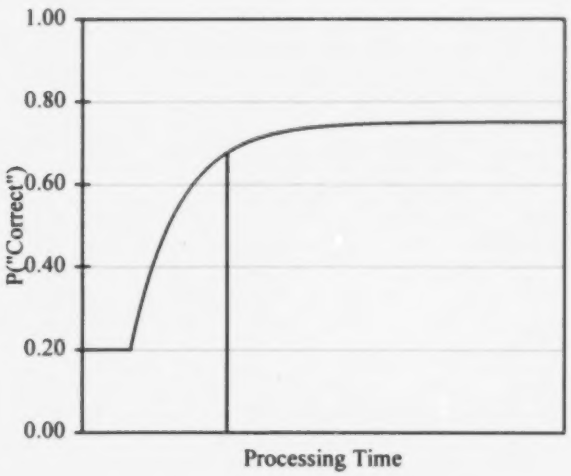

Figure 1. Theoretical function relating accuracy to processing time for a particular examinee responding to a particular five-choice item. Accuracy is an increasing function of processing time with a lower asymptote (chance) and an upper asymptote. The vertical line represents the examinee's chosen processing time and the resulting accuracy.

Although applying these empirical approaches in the testing context could be difficult and perhaps not useful, psychometric theories of response time may benefit from consideration of this within-subject relationship. To this end, Figure 1 suggests a possible theoretical description of the within-subject effect for a particular examinee using a particular solution strategy to respond to a particular five-choice item. The probability of responding correctly is assumed to be a monotonically increasing function of processing time with both a lower and upper asymptote.

Examinees are assumed to require some minimal amount of time in order to respond at above chance levels. This assumption is represented in Figure 1 by the lower asymptote at .20 (chance performance for a five-choice item). This minimal processing time could be considered strictly encoding time, but other processing stages (e.g., response initiation) might also be involved.

After the minimal processing has been accomplished, performance increases toward an upper asymptote. This upper asymptote represents the examinee's performance given an infinite amount of time. The examinee whose performance is depicted in Figure 1 has an asymptotic accuracy of approximately .75 for this particular item. Once an examinee has approached asymptotic accuracy, further processing would have little effect on performance. Thus, once our hypothetical examinee reaches an accuracy of .74, further processing can only increase accuracy by an additional .01.

Of course the shape of the function depicted in Figure 1 is purely arbitrary and only intended for descriptive purposes[1]. Still, the concepts of a minimal processing time, a monotonic increase in accuracy as a function of processing time, and an upper asymptote are useful constructs for considering response times in the testing context. Examinee, item, and strategy characteristics may be reflected in all three aspects of the speed-accuracy relationship.

If faced with a theoretical curve like the one depicted in Figure 1, cognitive psychologists might be interested in determining how characteristics of the curve (e.g., intercept, rate, and upper asymptote) could be explained in terms of underlying processes and operations. The general approach would be to produce empirical functions by requiring examinees to respond to items at specific response deadlines (Dosher, 1981; Reed, 1973). Curves based on items requiring different sets of cognitive processes or operations could then be compared.

[1] The deceleration depicted in Figure 1 suggests a rule of diminishing returns: once minimal processing is completed, each additional unit of processing time results in a smaller performance increase than the preceding unit. Laws of diminishing returns are common in cognitive psychology, but we would hesitate to make claims of the appropriateness of such an assumption in the testing context.

Psychometricians, on the other hand, might be more interested in individual differences. One particularly interesting issue would be the relationship between different aspects of the curves across examinees. For instance, is an examinee's minimal processing time or rate of increase in accuracy related to asymptotic accuracy? Unfortunately, producing the empirical curves of cognitive psychology is probably not feasible in the testing context and certainly not feasible with most operational data. Empirical speed-accuracy tradeoff functions of the sort depicted in Figure 1 require hundreds if not thousands of observations for each of a number of different response deadlines. Instead, psychometricians need to rely on more global measures of speed and accuracy, so the first task for theorists is to determine how speed-accuracy tradeoff functions are related to observable behavior in the testing context.

A Psychometric Approach to Speed-Accuracy Relationships

In a standard testing environment, examinees have direct control over strategy selection (e.g., heuristic versus algorithmic) and processing time but only indirect control over accuracy. Strategy selection affects the speed-accuracy relationship, and increased processing time generally leads to better performance.

In terms of the speed-accuracy tradeoff function depicted in Figure 1, strategy selection could affect the minimal processing time, the solution time, and the asymptotic accuracy. Examinees may select a strategy on the basis of any or all three characteristics of the speed-accuracy relationship. Of course, examinees differ in the number and type of strategies at their disposal, and examinee knowledge of these characteristics may be faulty, so strategy selection could be suboptimal.

Once a strategy has been selected, examinees can choose a response speed (i.e., the amount of time they are willing to expend on an item). This would be equivalent to selecting a point along the speed-accuracy tradeoff function depicted in Figure 1. The vertical line in the figure represents one possible choice; thus, this hypothetical examinee has decided to spend a certain amount of time on this particular item, and that amount of time results in approximately a .68 probability of responding correctly.

From this perspective, data from computer-administered tests provide a single data point for each of several speed-accuracy tradeoff functions for each examinee. The performance indices of CTT (percentage correct) and IRT (θ) are determined jointly by examinee ability, strategy selection, and choice of response speed. Furthermore, measures of response time alone do not provide enough information to relate an examinee's observed performance to potential (asymptotic) performance. With these important caveats in mind, the purpose of the present work is to explore the relationship between observed response speed and observed response accuracy in a standard testing context.

Thissen's (1983) Timed-Testing Model

Although the cognitive perspective discussed so far has its uses, a psychometric model of the speed-accuracy relationship may provide a more reasonable framework in the testing context. We will return to the issues developed in the previous sections in the Discussion.

Thissen (1983) offered an extension of Item Response Theory that not only captures examinee ability in terms of response accuracy but also accounts for response-time data in terms of item and examinee parameters. Additionally, the model allows for direct examination of the relationship between ability and speed.

Thissen's model consists of two related submodels: a response-accuracy submodel and a response-speed submodel. The response-accuracy model is the standard logistic IRT model:

$$P(x_{ij} = 1|\theta_i,\beta_j) = \frac{1}{1+e^{-z_{ij}}}, \quad (1)$$

where x_{ij} is 1 if Examinee i responds correctly to Item j, θ_i represents the ability of Examinee i, β_j represents item characteristics (perhaps vector-valued), and z_{ij} is a function of θ_i and β_j. Thissen used the two-parameter logistic (2PL) IRT model in which $\beta_j=(a_j, b_j)$ and $z_{ij}=1.702a_j(\theta_i-b_j)$. To help account for the lower-asymptote that arises in multiple-choice tests, we use the three-parameter logistic (3PL) in which $\beta_j=(a_j, b_j, c_j)$. The full 3PL model is given by:

$$P(x_{ij} = 1|\theta_i,\beta_j) = c_j + \frac{1-c_j}{1+e^{-z_{ij}}}, \quad (2)$$

where $z_{ij}=1.702a_j(\theta_i-b_j)$ as it does in the 2PL model. Equation 2 can be written in the form of Equation 1 by writing z_{ij} as:

$$z_{ij} = \ln[c_j + e^{1.702a_j(\theta_i-b_j)}] - \ln(1-c_j), \quad (3)$$

and this is the function we'll be using for z_{ij} throughout the remainder of the manuscript.

The response-speed model assumes lognormally distributed response-time distributions that are also a function of examinee and item parameters[2]:

$$\ln(t_{ij}) = \mu + \tau_i + s_j - \rho z_{ij} + \varepsilon, \quad (4)$$

where $\ln(t_{ij})$ is the natural logarithm of the time taken by Examinee i to respond to Item j, μ is a grand mean, τ_i is a slowness parameter for Examinee i, s_j is a slowness parameter for Item j, ρ is a regression coefficient representing the relationship between response time and relative item easiness (z_{ij}, given by Equation 3), and ε is a normal deviate with mean 0 and standard deviation σ. Notice that the normal distribution of ε causes t_{ij} to be lognormally distributed. This is

[2] Thissen's notation has been altered to avoid confusion with the 3PL response-accuracy model.

consistent with typically unimodal, positively skewed response-time distributions.

Thissen's model was applied to operational data from computer-based administrations of three large-scale standardized tests. Of particular interest are the estimated values of ρ and the relationship between θ_i and τ_j. The relationship between b_j and s_j was also examined because this is the item analog of the examinee analyses.

Method

Instrument

Data from computer-administered tests of verbal, quantitative, and reasoning skills were used for the present study. The items were administered non-adaptively; all examinees received the same items. Approximately 7,000 examinees completed all three tests.

Removal of Rapid Guessing

Aspects of the testing environment may affect both strategy selection and response speed. For example, restrictive time limits may encourage examinees to select less-accurate strategies over other strategies that require longer processing times. Such a situation is illustrated in Figure 2.

The examinee in Figure 2 has two solution strategies at her disposal: a heuristic strategy and an algorithmic strategy. If there are strict time limits, she may choose to minimize her processing time. This is indicated in the figure by assuming that she adopts the leftmost vertical line as her processing time. In this situation, she might want to use the heuristic strategy because it offers better performance at low processing times (note the relative heights of the curves where they cross the leftmost vertical line). If, on the other hand, she chooses to invest considerable time in this question (represented by the rightmost vertical line in the figure), she can increase her asymptotic accuracy by using the algorithmic strategy.

Strict time limits may cause examinees to emphasize processing speed to the exclusion of response accuracy. For example, as time begins to expire, examinees may begin to respond to items after only minimal processing. This is especially likely if no points are subtracted for incorrect responses. Schnipke and Scrams (in press) characterized such responding as rapid guessing. Such responding provides little if any information about examinee ability.

In order to minimize the effects of such responding, a technique suggested by Schnipke and Scrams (in press) was used to identify responses that probably reflect rapid guessing. Response-time distribution functions for all items are fit with two alternative models: a single-state model and a two-state model. The single-state model is a lognormal distribution function such as that predicted by the Timed-Testing model. This model should fit the unimodal, positively skewed distributions typical of response times.

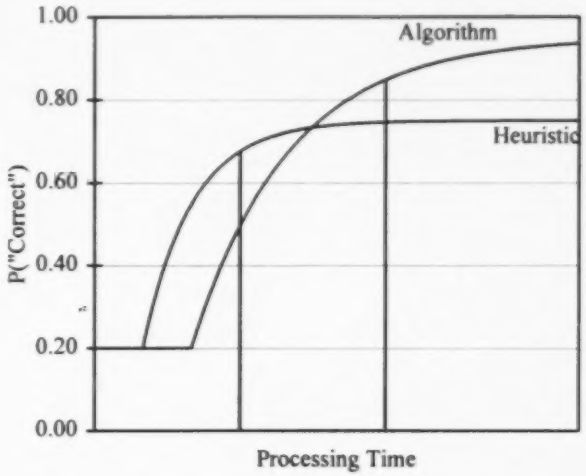

Figure 2. Comparison of two solution strategies in terms of their speed-accuracy tradeoff functions. The two vertical lines represent possible processing times.

If some examinees are engaging in rapid guessing, their response times will be unusually fast, and the resulting density function will be bimodal. The two-state model allows for examinees engaging in two different response strategies that result in different response-time distributions. This model can be used to deconvolve the observed response-time distribution into two component distributions: one for examinees actively trying to solve the item and one for examinees engaging in rapid guessing. Unfortunately, the model is statistical, so there is no way of knowing for certain which responses resulted from each strategy. Instead, responses are assigned a likelihood ratio that compares the estimated probability that the response reflects rapid guessing to the estimated probability that it reflects an active attempt to solve the item. If the likelihood ratio is greater than 1, the response is more likely to reflect rapid guessing than active solution, and the response is flagged as a potential rapid guess.

This technique was used with the present data, and the complete results for the reasoning test are reported by Schnipke and Scrams (in press). Rapid guessing was relatively uncommon on the verbal and quantitative tests; only a few items were affected. Rapid guessing was found only on one item type on the verbal test (11 out of 38 total items) and for the last five of the 30 items on the quantitative test. Although relatively few examinees were identified as engaging in rapid guessing, all of these items were removed from the present analyses.

Unfortunately, rapid guessing was more widespread on the reasoning test (affecting over half of the 25 items). Removing all affected items would have severely reduced the number of usable items, so the affected items were retained, but all responses identified as likely to reflect rapid guessing

(a likelihood ratio greater than 1) were treated as unseen. This is a suboptimal solution given that some of the remaining responses probably reflect rapid guessing, and some of the responses that were treated as unseen were probably not rapid guesses.

Parameter Estimation

Thissen (1983) provides a joint maximum likelihood equation for fitting parameters of the 2PL version of his Timed-Testing model, but a simpler procedure was used for the present analyses. Equation 4 can be considered an analysis of covariance with a constant (μ), two factors (examinee and item), and one covariate (z_{ij}), and we used this conceptualization to fit the model to the present data.

The response-accuracy and response-speed submodels were estimated separately. First, standard 3PL IRT parameters were estimated using BILOG (Mislevy & Bock, 1990). These parameters were used to estimate z_{ij}'s. These estimates were treated as true parameters for estimation of the response-speed parameters. Standard analysis of covariance techniques were used to estimate the parameters of the response-speed submodel. Although the full set of 7,218 examinees was used to estimate the IRT parameters, a subset of 1,000 examinees was selected for the analysis of covariance procedure for computational reasons.

The joint maximum likelihood approach might provide a better overall fit by sacrificing the fit of both submodels. In our case, we maximize the fit of the response-accuracy model, so our response-speed parameters may be suboptimal. This, however, is a simpler procedure, we assume that the error is non-systematic, and simulation results suggested that the approach performed very well with sample sizes similar to the ones used in the present work.

Results

The relationship between speed and accuracy was similar for the verbal and quantitative tests but different for the reasoning test. Item differences tended to account for the majority of explained variability in log response times, and relative item easiness (z_{ij}) accounted for very little variability. High-ability examinees tended to have higher slowness indices than low-ability examinees on the verbal and quantitative tests, but speed and accuracy were unrelated on the reasoning test. Item difficulty and item slowness were unrelated on all three measures. Specific results are presented separately for each test.

Verbal Test

For the verbal test, the analysis of covariance model given in Equation 4 accounted for 54.43% of the variability in log response times. Of the total variability, 33.06% was explained uniquely by item differences, 8.80% was explained uniquely by examinee differences, and 1.05% was explained uniquely by relative item easiness (z_{ij}). The remaining explained variability (11.52%) was shared among two or more predictors.

The estimated value of ρ, the regression coefficient relating log response time to relative item easiness, was .19, and the estimate of μ, the grand mean log response time, was 2.81 log seconds (16.62 seconds).

Examinee ability (θ) and examinee slowness (τ) were correlated (r^2=.39), but the direction of the relationship was somewhat counter to expectation. As shown in Figure 3, examinee ability and examinee slowness were positively correlated; higher ability examinees had higher slowness indices.

Item difficulty (b) and item slowness (s) were unrelated (r^2=.02).

Quantitative Test

For the quantitative test, the analysis of covariance model accounted for 43.39% of the variability in log response times. Of the total variability, 27.53% was explained uniquely by item differences, 11.52% was explained uniquely by examinee differences, and 0.72% was explained uniquely by relative item easiness (z_{ij}). The remaining explained variability (3.62%) was shared among two or more predictors.

The estimated value of ρ was .13, and the estimate of μ was 4.25 log seconds (69.96 seconds).

Examinee ability (θ) and examinee slowness (τ) were correlated (r^2=.33), and the direction of the relationship was consistent with the verbal test results. As shown in Figure 4, examinee ability and examinee slowness were positively correlated.

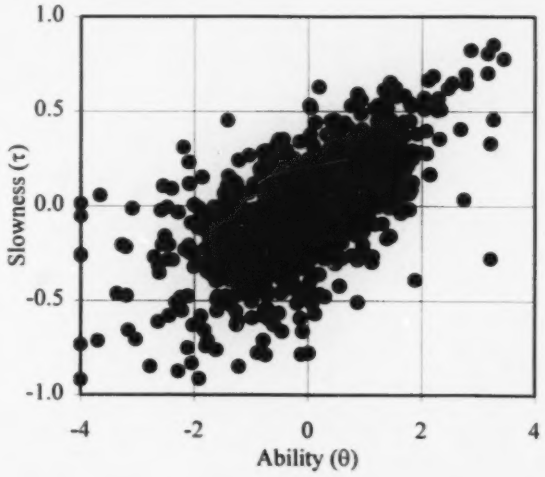

Figure 3. Scatterplot showing the relationship between examinee slowness (τ) and examinee ability (θ) for the verbal test.

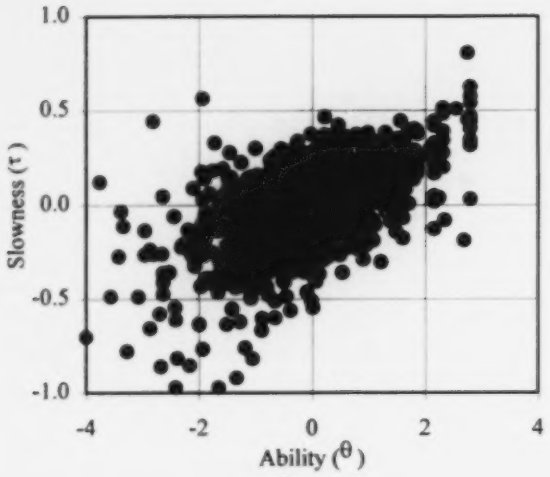

Figure 4. Scatterplot showing the relationship between examinee slowness (τ) and examinee ability (θ) for the quantitative test

Item difficulty (*b*) and item slowness (*s*) were unrelated (r^2=.02).

Reasoning Test

For the reasoning test, the analysis of covariance model accounted for 40.46% of the variability in log response times. Of the total variability, 33.61% was explained uniquely by item differences, 6.31% was explained uniquely by examinee differences, and 0.02% was explained uniquely by relative item easiness (z_{ij}). Relatively little variability (0.52%) was shared among two or more predictors.

The estimated value of ρ was -.02, and the estimate of μ was 4.18 log seconds (65.46 seconds).

Examinee ability (θ) and examinee slowness (τ) were unrelated (r^2=.00), but a scatterplot is provided as Figure 5 for comparison with the results for the other tests.

Item difficulty (*b*) and item slowness (*s*) were also unrelated (r^2=.00).

Alternative Analyses

Interpretation of the examinee ability-slowness and item difficulty-slowness relationships is complicated because relative item easiness (z_{ij}) is a function of examinee ability and item characteristics. Effects on item response time due to examinee ability (θ) and item difficulty (*b*) may be incorporated into the slowness indices (τ and *s*). This could cause the effects of θ and *b* to be masked (or even reversed).

To examine this possibility for the present data, an additional regression analysis was undertaken. For each test, log response time was regressed on examinee ability (θ), item difficulty (*b*), and relative item easiness (z_{ij}). If fitting the submodels separately masked the effects of examinee ability and item difficulty, these predictors should account for a significant proportion of variability in log response times.

For the verbal test, examinee ability, item difficulty, and relative item easiness accounted for a total of 12.51% of the variability in log response times. Most of this variability was shared among multiple predictors (9.52% of total variability), and an additional 2.14% was explained uniquely by relative item easiness. Very little variability was explained uniquely by examinee ability (0.84%) or item difficulty (0.00%).

Even less variability was explained in the analyses of the quantitative (6.21%) and reasoning (0.52%) tests. Most of the variability that was explained in the analysis of the quantitative test was shared among multiple predictors (4.24% of total variability). Examinee ability, item difficulty, and relative item easiness uniquely accounted for 0.58%, 0.65%, and 0.74% of total variability, respectively.

Results for the reasoning test were similar to those for the quantitative test. Most of the explained variability was shared among multiple predictors (0.36% of total variability). Examinee ability, item difficulty, and relative item easiness uniquely accounted for 0.03%, .14%, and 0.00%, respectively.

Re-Examining Relative Item Easiness

Relative item easiness (z_{ij}) was a poor predictor in both the model-driven and alternative analyses, but the 3PL model used in the present work incorporates four factors into determining relative item easiness: examinee ability (θ), item difficulty (*b*), item discrimination (*a*), and lower asymptotic performance (*c*). An additional set of regression analyses was performed using a simpler measure of relative item easiness: θ_r-b_i. This is the logit from the one-parameter logistic (Rasch) IRT model.

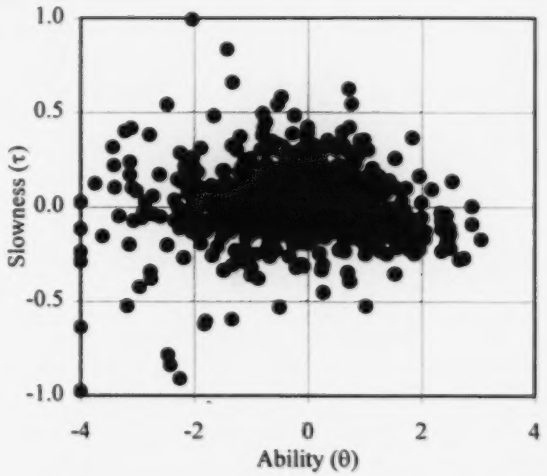

Figure 5. Scatterplot showing the relationship between examinee slowness (τ) and examinee ability (θ) for the reasoning test.

For each test, log response time was regressed on the alternative measure of relative item easiness (θ_i-b_j), but little variability was explained: 8.50%, 3.19%, 0.48% for the verbal, quantitative, and reasoning tests, respectively. Adding a quadratic term to the model increased the explained variability by a modest amount to 11.19%, 5.55%, and 0.51% for the verbal, quantitative, and reasoning tests, respectively.

Discussion

A variety of analyses were conducted to examine the relationship between response speed and response accuracy in the context of large-scale, standardized test administrations. Overall, relative item easiness accounted for only a small proportion of variability in response times, and neither examinee ability (θ) nor item difficulty (b) performed much better.

The item and examinee indices derived from Thissen's Timed-Testing model (1983) demonstrated the strongest relationships. For the verbal and quantitative tests, examinee ability was positively correlated with examinee slowness. In light of the other results, these correlations probably reflect the influence of correcting the slowness indices for the negative effect of relative item easiness. In other words, the two results tend to cancel one another, so they may represent effects of sampling error in parameter estimation.

On the surface, these results are incongruent with the belief that higher ability examinees also process information more quickly, but the cognitive perspective illustrated in Figures 1 and 2 suggests an important caveat. In standard administrations, examinees select the amount of time they are willing to spend on items, and their selections are likely to be influenced by characteristics of the testing situation.

The test analyzed in the present work was administered with a time limit. Not all examinees used all the time available, and not all examinees completed all items, but the time limit may have encouraged most examinees to adopt similar pacing strategies. This would reduce the amount of variability explained by examinee differences. This is consistent with the results of the model-based analyses in which examinee differences uniquely accounted for only 8.80% (verbal test), 11.52% (quantitative test), and 6.31% (reasoning test) of variability in log response times. This lack of variability across examinees could result in underestimation of the effects of examinee ability.

In terms of the theoretical description illustrated in Figures 1 and 2, high-scoring examinees may be performing near asymptotic accuracy. This might allow such examinees to increase their response speed with only minor reductions in observed accuracy.

An interesting question concerns the placement of low-scoring examinees on their respective speed-accuracy tradeoff functions. Perhaps, low-scoring examinees are also performing near asymptotic accuracy. In this case, high- and low-scoring examinees differ in asymptotic accuracy. Alternatively, low-scoring examinees may be responding well below their asymptotic accuracy because time limits prevent them from spending the time necessary to perform at their optimal level. In this case, high- and low-scoring examinees may differ primarily in processing speed (either the rate of increase in accuracy or in the minimum processing time required). Whether high- and low-scoring examinees differ in asymptotic accuracy or in processing speed has serious consequences for the interpretation of test speededness.

Speededness

Time limits are often used as an administrative convenience, and the effects of speededness are often seen as unrelated to what the test is measuring. Many such tests are intended to be power tests, in which case interest is on asymptotic accuracy, the accuracy examinees could achieve if there were no time limits. When there are time limits on the test, examinees may be required to answer faster, at least on some items, thus lowering their accuracy on those items.

If examinees only differ in their asymptotic accuracy (and not in their minimum processing time or rate of rise), then time limits won't matter because the ordering of examinees will be the same (their accuracy functions will not cross). If examinees differ in their processing rate (their minimum processing time or rate of rise), the accuracy functions may cross or at least come together. In this case, speededness is an issue because examinee ordering can change.

Incorporating Speed into Ability Estimation

The present results could be interpreted as evidence for the potential usefulness of incorporating speed indices into ability estimation. For the present data, speed and accuracy are at least separable if not orthogonal. However, scoring examinees on the basis of their speed as well as their accuracy introduces several potential problems.

The use of time limits already requires examinees to select a solution strategy and processing speed with the hopes of maximizing their observed performance. If examinees are scored on their response speed as well, their choice of strategy and speed will be even more complicated.

Additionally, if response speed is an explicit part of the score, the validity and interpretation of the score needs to be established. Thus we are not recommending that speed be incorporated into ability estimation before extensive research establishes how to interpret and use such scores.

Conclusions

The availability of response times from operational tests suggests a number of useful avenues of research and allows a more detailed examination of the relationship between speed and accuracy in the testing context than has previously been possible. However, this relationship is likely to be very complicated, and significant theoretical work is necessary to

accommodate these complications. The within-examinee conceptualization borrowed from cognitive psychology and illustrated in Figures 1 and 2 may be a good beginning or at least may encourage psychometricians to consider some of the difficulties inherent in interpreting response-time data.

References

Dosher, B. A. (1981). The effects of delay and interference: A speed-accuracy study. *Cognitive Psychology, 13*, 551-582.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.

Luce, R. D. (1986). Response times: Their role in inferring elementary mental organization. New York: Oxford University Press.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models [Computer software and manual]. Chicago: Scientific Software, Inc.

Reed, A. V. (1973). Speed-accuracy tradeoff in recognition memory. *Science, 181*, 574-576.

Schnipke, D. L., & Scrams, D. J. (in press). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika*, 2, 249-254.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University.